MINI PROJECT REPORT - SLR + SLC + USL

ANALYSIS OF ZOMATO FOOD DELIVERY CASE STUDY

GROUP 3

CONTENTS

Topics

TEAM MEMBERS

Aditya Aryan

Bhavani Kanaparthi

Naman Goswami

Neha Mondal

INTRODUCTION TO THE PROBLEM STATEMENT

Zomato is an Indian multinational restaurant aggregator and food delivery company founded by Deepinder Goyal and Pankaj Chaddha in 2008. Zomato provides information, menus, and user reviews of restaurants as well as food delivery options from partner restaurants in select cities.

Restaurants from all over the world can be found in Bengaluru. From the United States to Japan, Russia to Antarctica, you get all types of cuisines here. Delivery, Dine-out, Pubs, Bars, Drinks, Buffet, Desserts, you name it, and Bengaluru has it. Currently, it stands at approximately 12,000 restaurants. And new restaurants are opening every day. However, it has become difficult for them to compete with already established restaurants. The key issues that continue to pose a challenge to them include high real estate costs, rising food costs, shortage of quality manpower, fragmented supply chain, and over-licensing.

Section A

From the data we have been given we have to predict the cost per two customers for one time so that the newly started restaurants and upcoming restaurants will be well prepared how the restaurant should invest in improving the ambiance and all other stuff to attract the customers. And to help new and upcoming restaurants by letting them know the various reasons that customers look for and build a model which can predict the cost for two people.

Section B

Now Zomato has been observing the orders happening online and offline. Due to offline orders, Zomato is not able to attract customers with diverse items and offers, and the user subscription is also getting low. Zomato wants to know whether the customer would order the orders online or offline so that it can take further strategies to improve the online order. So, with the data provided we must classify the orders that have been ordered online and offline and identify the patterns that lead to orders online orders as well as offline.

TECHNOLOGIES USED

While the theoretical aspects and knowledge and experience gained from a project are significant, no Data Science project can be finished without the use of technology. It is essential to fulfil the asks of the problem statement. Hence, it is important we address the non-theoretical sides behind the outcomes of this project.

We have primarily come across two types of technologies throughout the creation of our project. One of them is the programming language we have used for various calculations, data cleaning and visualization etc. while the other is the platform which provided us with the facility of making use of the said programming language. They are listed as follows:

- ✓ **Python (Programming Language)**

  Python is a high-level, interpreted and general-purpose programming language. It is often known as the "Swiss Army Knife" of programming languages. Commended for its simplicity, readability and versatility, it uses a very simple and easy to understand syntax which makes it a perfect choice for beginners to programming.

- ✓ **Jupyter Notebook (Integrated Development Environment)**

  Jupyter Notebook is a web-based programming environment which is open-source. It provides its users the ability to create and edit programs, graphs and comments all in a single document. It is a very popular tool used by data scientists and mentors of data analysis and machine learning. It is used to code in various programming languages, including Python, R, Julia, and many others.

SKILLS USED AND DEVELOPED

The skills we have used are as follows:

- ✓ **Programming:** Basic Python programming has helped us in using libraries, looping etc. throughout the timeline of the project.
- ✓ **Inferencing:** In both the sections, we have made numerous inferences by looking at numbers and graphs.
- ✓ **Data Pre-processing:** We have performed outlier detection, dropping and conversion of columns, null removal and imputation on the data provided.
- ✓ **Data Visualization:** Various graphs and plots have been created from our knowledge of visualization to dig deeper into the data for information.
- ✓ **Statistical Analysis:** We have used statistical concepts for finding out probabilities and testing hypotheses and claims.
- ✓ **Domain Knowledge:** For both the sections, domain knowledge has been used to make significant decisions at times.
- ✓ **Exploratory Data Analysis:** To analyse the data for hidden information and to identify patterns and insights from the data, we have made use of the concepts of EDA.
- ✓ **Problem-Solving:** While we were mostly told what to do to achieve results, it took us to solve undefined problems at instances.
- ✓ **Supervised Learning Regression:** As the target variable in first section was numeric, regression was needed to be done.
- ✓ **Supervised Learning Classification:** As the target variable in second section was categorical, classification was needed to be done.
- ✓ **Unsupervised Learning:** In order to know additional information about the data, we performed clustering in the second section.

Lastly, the application of above-mentioned skills has helped us in their development on a certain level.

DATA DESCRIPTION

The data set we are provided with is filled with orders from different restaurants with data such as URL, address and name of the restaurant, type of order, table booking etc.

There are 51,717 records and a total of 17 columns in the data. 16 of the columns are originally categorical while the remaining one is numerical. 19,720 records were found to be duplicated, and several columns have missing values going up to a maximum percentage of 19.44. The data dictionary is as follows:

No.	Column Name	Description	Data Type
1	URL	Website of Zomato for each restaurant	Object
2	Address	Address of the restaurant	Object
3	Name	Name of the restaurant	Object
4	Online Order	Indicates if the customer ordered the menu online or not	Object
5	Book table	Indicates if the customer has booked the table or not	Object
6	Rate	Rating of the restaurant that has been given by the customers	Numerical
7	Votes	The number of votes given by the customers to the restaurant	Numerical
8	Phone	Contact number of the restaurant.	Object
9	Location	The city name where the restaurant is located	Object
10	Rest Type	The type of restaurant	Object
11	Dish liked	Dishes liked by the customer from the restaurant	Object
12	Cuisines	The cuisines that are prepared by the restaurant	Object
13	Approx Cost for two people	The approximate cost for 2 people	Numerical
14	Reviews list	The reviews made by the customers on the restaurant	Object
15	Menu Item	The menu items that are usually available at the restaurant	Object
16	Listed in (type)	The type of the meal	Object
17	Listed in (city)	The neighbourhood in which the restaurant is listed	Object

DATA COLLECTION AND CLEANING

Our data was taken from the “zomato.csv” file which was provided to us.

We saw that the columns “rate” and ‘approx_cost(for two people)’ had symbols and spaces etc. which made them have an object data type initially and thus we removed them and converted the values in them in float.

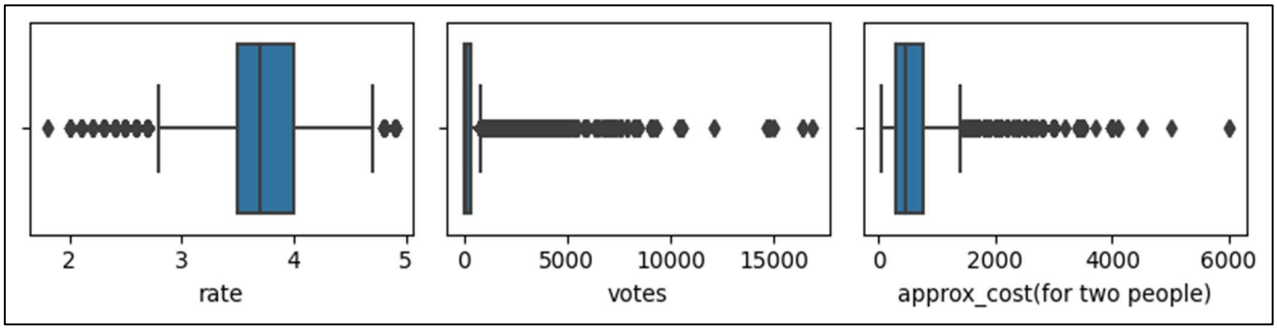
```
# cleaning of 'rate'

# function replaces all 'NEW' and '-' values with null so that we can handle them easily
# function also removes the '/' part to change the value into float
def clean_rate(val):
    if val == "NEW" or val == "-":
        return np.nan
    else:
        val = str(val).split('/')
        val = val[0] # takes only the part before '/'
        return float(val)
d.rate = d.rate.apply(clean_rate)
d.rate.unique()
```

```
array([4.1, 3.8, 3.7, 3.6, 4.6, 4. , 4.2, 3.9, 3.1, 3. , 3.2, 3.3, 2.8,
       4.4, 4.3, nan, 2.9, 3.5, 2.6, 3.4, 4.5, 2.5, 2.7, 4.7, 2.4, 2.2,
       2.3, 4.8, 4.9, 2.1, 2. , 1.8])
```

‘url’ and ‘phone’ variables were not important for our analysis. Since we already had the data about the area of the restaurant in the ‘location’ variable, we did not require ‘address’ and ‘listed in(city)’. Since there were multiple values present in one entry of ‘reviews_list’, ‘menu_item’, ‘dish_liked’ and ‘cuisines’ and thus there were a lot of unique values, the cardinality of these variables was high. Thus, we dropped these columns.

‘rest_type’ and ‘approx_cost(for two people)’ had a very low percentage of missing values; thus, we dropped them. ‘rate’ had negative outliers and thus, we imputed its missing values with its median.

We saw that all the numeric variables had outliers. ‘votes’ and ‘approx_cost(for two people)’ had too many outliers and thus we kept them unchanged to prevent loss of data. We capped the outliers of 'rate' using IQR method.



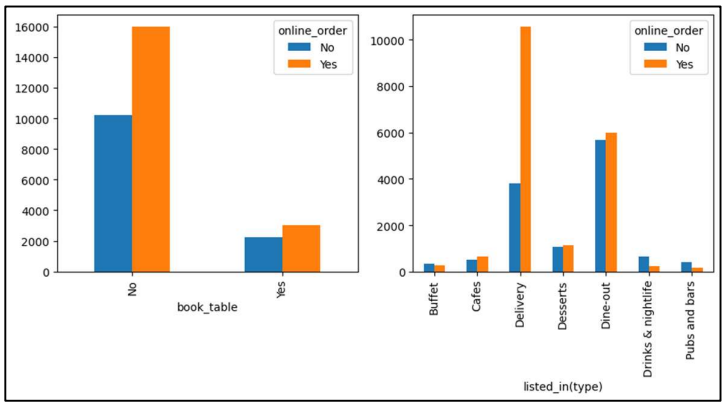
PROBLEM SOLVING STEPS

Section A

- ✓ In this case, our target variable was 'approx_cost(for two people)'.
- ✓ After cleaning the data as per our requirement, we did univariate and bivariate analysis to find out patterns and significant features.
- ✓ We did some hypothesis testing to verify the significance of different attributes on our target variable. For attributes "online_order" and "booked_table" we performed T-test to check that all the different categories in the columns have different mean costs and concluded that the target depends on them. For columns 'rate' and 'votes', we performed Pearson's Test of Correlation with the target variable.
- ✓ We encoded our categorical variables using various techniques.
- ✓ We split our data into 30% test and 70% train set.
- ✓ We built our base model using Ordinary Least Squares method.
- ✓ We built Random Forest as our final model which gave us desirable results.

Section B

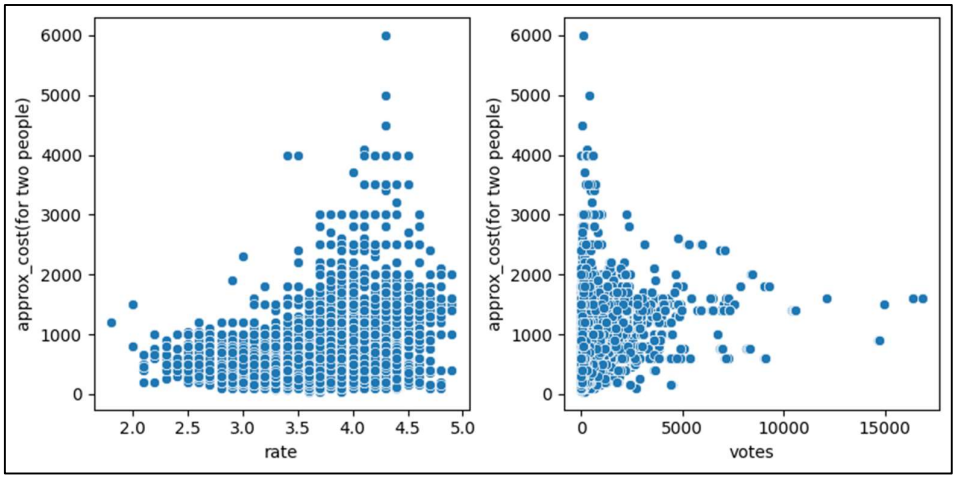
- ✓ We used the same data that was cleaned. For this case, our target variable was the column "online_order". We observed slight imbalance in it but we could continue with our analysis without making any changes.
- ✓ We performed bivariate analysis of our target variable against the categorical and numerical variables.
- ✓ Then we took the numerical columns and scaled them using z-score to perform clustering.
- ✓ Using elbow plot, we decided the optimal no. of clusters for the K-means clusters for our data, which came to be 5. We categorized the data into five clusters.



- ✓ We interpreted different clusters and checked whether they were dominated by online or offline orders.
- ✓ We performed Chi-square Test of Independence of target variable on 'listed_in(type)''and our new variable 'cluster'. Confirmed that the target variable does depend on them.
- ✓ We encoded our categorical variables according to the new target variable.
- ✓ We split our data into 30% test and 70% train set.
- ✓ We trained our base model using Logistic Regression.
- ✓ We fed our data in different models like Logit model, K-Nearest Neighbours model, Gaussian Naive Bayes model, Decision Tree and Random Forest model.
- ✓ The Random Forest model was chosen as our final model based on its performance. We performed the prediction on test set.

TAKEAWAYS AND CONCLUSIONS

1. The average rating of all the restaurants is 3.7.
2. The average cost for two people in any restaurant is 555.43.
3. The rating of restaurants ranges from a lowest of 1.8 to a highest of 4.9.
4. The approximate cost for two people ranges from the lowest of 40 to a highest of 6000.
5. 'Onesta' and 'Cafe Coffee Day' are two of the most common restaurants.
6. BTM location has the highest number of restaurants.
7. 'Casual Dining' is the most common type of restaurants after 'Quick Bites'.
8. Offline orders are not much less than online orders.
9. Delivery and dine-out type of restaurants are much more than any other type.
10. Most of the restaurants are rated between 3 and 4.5.
11. Restaurants with zero votes are common.
12. The approximate cost for two people does not go beyond 2000 in most of the restaurants.
13. Orders with tables booked have approximately higher costs.
14. 'Buffet', 'Drinks & nightlife' and 'Pubs and bars' type of restaurants have higher approximate cost.
15. The approximate cost is higher for higher rated restaurants.
16. Restaurants in which the approximate cost is high have less votes.



Conclusions

1. Restaurants should try to promote offline orders as much as possible by locating the restaurant near residential areas and having better telephone service.
2. They should invest in abundant seating so that customers do not feel the need of booking tables.
3. Enabling delivery option and providing more variety in desserts and dining food items can decrease costs.
4. If a customer wants to book tables, the order is more likely to be online than offline.
5. Higher rated restaurants have a higher chance of online order.
6. Orders from delivery type restaurants are much more likely to be online than other types.
7. Buffet restaurants, pubs and bars generally have a higher chance of offline orders than online.
8. In the future, we can expect an increase in online orders as the online delivery apps grow.

FUTURE STEPS

- We can use libraries like Lime in Python to explain the predictions of our trained models in detail.
- Model deployment packages like Streamlit in Python can be used to deploy the model and observe efficient results.
- To share our findings with clients, we can create dashboards in visualization tools like Tableau or Power BI.